

SORT: Second-Order Response Transform for Visual Recognition

Yan Wang¹, Lingxi Xie², Chenxi Liu², Ya Zhang¹, Wenjun Zhang¹, Alan Yuille²

¹School of Electronic Info. and Electrical Engi., Shanghai Jiao Tong University, Shanghai, China

tiffany940107@gmail.com yazhang@sjtu.edu.cn zhangwenjun@sjtu.edu.cn

²Center for Imaging Science, The Johns Hopkins University, Baltimore, MD, USA

198808xc@gmail.com cxliu@jhu.edu alan.l.yuille@gmail.com

Abstract

In this paper, we reveal the importance and benefits of introducing second-order operations into deep neural networks. We propose a novel approach named *Second-Order Response Transform (SORT)*, which appends element-wise product transform to the linear sum of a two-branch network module. A direct advantage of *SORT* is to facilitate cross-branch response propagation, so that each branch can update its weights based on the current status of the other branch. Moreover, *SORT* augments the family of transform operations and increases the nonlinearity of the network, making it possible to learn flexible functions to fit the complicated distribution of feature space. *SORT* can be applied to a wide range of network architectures, including a branched variant of a chain-styled network and a residual network, with very light-weighted modifications. We observe consistent accuracy gain on both small (CIFAR10, CIFAR100 and SVHN) and big (ILSVRC2012) datasets. In addition, *SORT* is very efficient, as the extra computation overhead is less than 5%.

1. Introduction

Deep neural networks [24][43][46][13] have become the state-of-the-art systems which are widely used in visual recognition. Supported by large-scale labeled datasets such as **ImageNet** [5] and powerful computational resources like modern GPUs, it is possible to train a hierarchical structure to capture different levels of visual patterns. A pre-trained deep network is also capable of generating transferrable features for different vision tasks such as image classification [6], instance retrieval [40] and object detection [10], or fine-tuned to deal with a wide range of challenges, including object detection [41], semantic segmentation [34][2], boundary detection [54], etc.

The past years have witnessed an evolution in designing efficient network architectures, in which the chain-styled modules have been extended to multi-path modules [46]

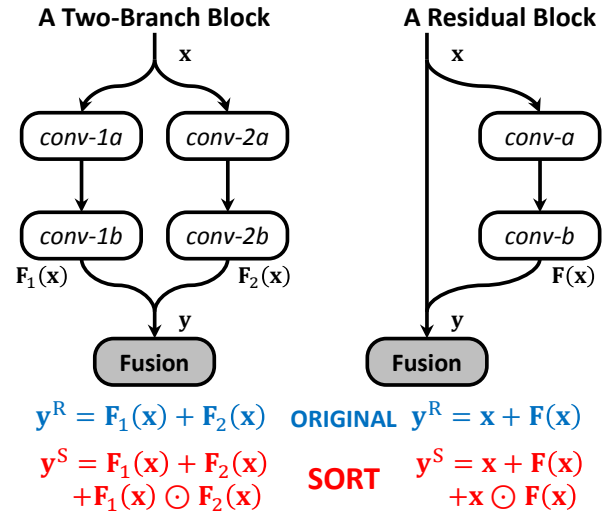


Figure 1. Two types of modules and the corresponding SORT operations. Left: in a two-branch convolutional block, the two-way outputs, $F_1(x)$ and $F_2(x)$, are combined with a second-order transform $F_1(x) + F_2(x) + F_1(x) \odot F_2(x)$. Right: in a residual-learning building block [13], we can also modify the fusion function from $x + F(x)$ to $x + F(x) + x \odot F(x)$. In both cases, \odot denotes the element-wise product operation.

or residual modules [13]. Meanwhile, highway inter-layer connections are verified helpful in training very deep networks [44]. In the previous literatures, highway and/or residual connections are fused in a linear manner, in which the neural responses of two branches are element-wise summed up as the output. This limits the ability of a deep network to fit the complicated distribution of feature space, as nonlinearity forms the main contribution to the network capacity [21]. This motivates us to consider higher-order transform operations.

In this paper, we propose *Second-Order Response Transform (SORT)*, an efficient approach that applies to a wide range of visual recognition tasks. The core idea of *SORT* is to append a dyadic second-order operation, say element-

wise product, to the original linear sum of two-branch vectors. This modification, as shown in Figure 1, brings two-fold benefits. First, SORT facilitates *cross-branch* information propagation, which rewards consistent responses in forward-propagation, and enables each branch to update its weights based on the current status of the other branch in back-propagation. Second, the nonlinearity of the module becomes stronger, which allows the network to fit more complicated feature distribution. In addition, adding such operations is very cheap, as it requires less than 5% extra time, and no extra memory consumptions. We apply SORT to both deep chain-styled networks and deep residual networks, and verify consistent accuracy gain over some popular visual recognition datasets, including **CIFAR10**, **CIFAR100**, **SVHN** and **ILSVRC2012**. SORT also generates more effective deep features to boost the transfer learning performance.

The remainder of this paper is organized as follows. Section 2 briefly reviews related work. Section 3 illustrates the second-order response transform algorithm and analyzes its benefits. Experiments are shown in Section 4, and conclusions are drawn in Section 5.

2. Related Work

2.1. Convolutional Neural Networks

The Convolutional Neural Network (CNN) is a hierarchical model for visual recognition. It is based on the observation that a deep network with enough neurons is able to fit any complicated data distribution. In past years, neural networks were shown effective for simple recognition tasks [27]. More recently, the availability of large-scale training data (*e.g.*, ImageNet [5]) and powerful GPUs make it possible to train deep architectures [24] which significantly outperform the conventional Bag-of-Visual-Words [25][48][39] and deformable part models [8]. A CNN is composed of several stacked layers. In each of them, responses from the previous layer are convoluted with a filter bank and activated by a differentiable non-linearity. Hence, a CNN can be considered as a composite function, which is trained by back-propagating error signals defined by the difference between supervision and prediction at the top layer. Recently, efficient methods were proposed to help CNNs converge faster and prevent over-fitting, such as ReLU activation [37], Dropout [15], DisturbLabel [53], batch normalization [19] and varying network depth in training [18]. It is believed that deeper networks have stronger ability of visual recognition [43][46][13], but at the same time, deeper networks are often more difficult to be trained efficiently [45].

An intriguing property of the CNN lies in its transfer ability. The intermediate responses of CNNs can be used as effective image descriptors [6], and widely applied to

various types of vision applications including image classification [22][52], image retrieval [40][50] and object detection [10]. Also, the pre-trained networks on a large-scale dataset can be fine-tuned to deal with other tasks, including object detection [41], semantic segmentation [2], boundary detection [54], *etc.*

2.2. Multi-Branch Network Connections

Beyond the conventional chain-styled networks [43], it is observed that adding some sideway connections can increase the representation ability of the network. Typical examples include the inception module [46], in which neural response generated by different kernels are concatenated to convey multi-scale visual information. Meanwhile, the benefit identity mapping [14] motivates researchers to explore networks with residual connections [13][56][17]. These efforts can be explained as the pursuit of building highway connections to prevent gradient vanishing and/or explosion in training very deep networks [44][45].

Another family of multi-branch networks follow the bilinear CNN model [32], which constructs two separate streams to model the co-occurrence of local features. Formulated as the outer-product of two vectors, it requires a larger number of parameters and more computational resources than the conventional models to be trained. An alternative approach is proposed to factorize bilinear models [30] for visual recognition, which largely decreases the number of trainable parameters.

All the multi-branch structures are followed by a module to fuse different sources of features. This can be done by linearly summing them up [13], concatenating them [46], and using a bilinear [32] or recurrent [45] transform. In this work, we present an extremely simple and efficient approach to enable effective feature ensemble, which involves introducing a second-order term to apply nonlinear transform in neural responses. This, as shown later, increases the network capacity and leads to improved effectiveness in network training.

3. Second-Order Response Transform

3.1. Formulation

Let \mathbf{x} be a set of neural responses at a given layer of a deep neural network. In practice, \mathbf{x} often appears as a 3D volume. In a two-branch network structure, \mathbf{x} is fed into two individual modules with different parameters, and two intermediate data cubes are obtained. We denote them as $\mathbf{F}_1(\mathbf{x}; \theta_1)$ and $\mathbf{F}_2(\mathbf{x}; \theta_2)$, respectively. In the cases without ambiguity, we write $\mathbf{F}_1(\mathbf{x})$ and $\mathbf{F}_2(\mathbf{x})$ in short. Most often, $\mathbf{F}_1(\mathbf{x})$ and $\mathbf{F}_2(\mathbf{x})$ are of the same dimensionality, and an element-wise operation is used to summarize them into the output set of responses \mathbf{y} .

There are some existing examples of two-branch net-

works, such as the Maxout network [11] and the deep residual network [13]. In Maxout, $\mathbf{F}_1(\mathbf{x})$ and $\mathbf{F}_2(\mathbf{x})$ are generated by two individual convolutional layers, *i.e.*, $\mathbf{F}_m(\mathbf{x}) = \sigma[\boldsymbol{\theta}_m^\top \mathbf{x}]$ for $m = 1, 2$, where $\boldsymbol{\theta}_m$ is the m -th convolutional matrix, $\sigma[\cdot]$ is the activation function, and an element-wise max operation is performed to fuse them: $\mathbf{y}^M = \max\{\mathbf{F}_1(\mathbf{x}), \mathbf{F}_2(\mathbf{x})\}$. In a residual module, $\mathbf{F}_1(\mathbf{x})$ is simply set as an identity mapping (*i.e.*, \mathbf{x} itself), and $\mathbf{F}_2(\mathbf{x})$ is defined as \mathbf{x} followed by two convolutional operations, *i.e.*, $\mathbf{F}_2(\mathbf{x}) = \boldsymbol{\theta}_2^\top \sigma[\boldsymbol{\theta}_2^\top \mathbf{x}]$, and the fusion is performed as linear sum: $\mathbf{y}^R = \mathbf{F}_1(\mathbf{x}) + \mathbf{F}_2(\mathbf{x})$. Other examples of multi-branch network include [49].

The core idea of SORT is extremely simple. We append a second-order term, *i.e.* element-wise product, to the linear term, leading to a new fusion strategy:

$$\mathbf{y}^S = \mathbf{F}_1(\mathbf{x}) + \mathbf{F}_2(\mathbf{x}) + \mathbf{F}_1(\mathbf{x}) \odot \mathbf{F}_2(\mathbf{x}). \quad (1)$$

Here, \odot denotes the element-wise product of $\mathbf{F}_1(\mathbf{x})$ and $\mathbf{F}_2(\mathbf{x})$. The gradient of \mathbf{y}^S over either \mathbf{x} or $\boldsymbol{\theta}_m$ ($m = 1, 2$) is straightforward. Note that this modification is very simple yet light-weighted. Based on a specifically implemented layer in popular deep learning tools such as CAFFE [22], SORT requires less than 5% additional time in training and testing, meanwhile no extra memory is used.

SORT can be applied to a wide range of network architectures, even if the original structure does not have branches. In this case, we need to modify each of the original convolutional layers, *i.e.*, $\mathbf{y}^O = \sigma[\boldsymbol{\theta}^\top \mathbf{x}]$. We construct two symmetric branches $\mathbf{F}_1(\mathbf{x})$ and $\mathbf{F}_2(\mathbf{x})$, in which the m -th branch is defined as $\mathbf{F}_m(\mathbf{x}) = \sigma[\boldsymbol{\theta}_m'^\top \sigma[\boldsymbol{\theta}_m^\top \mathbf{x}]]$. Then, we perform element-wise fusion (1) beyond $\mathbf{F}_1(\mathbf{x})$ and $\mathbf{F}_2(\mathbf{x})$. Following the idea to reduce the number of parameters [43], we shrink the receptive field size of each convolutional kernel in $\boldsymbol{\theta}_m$ from $k \times k$ to $\lfloor \frac{1}{2}(k+1) \rfloor \times \lfloor \frac{1}{2}(k+1) \rfloor$. With two cascaded convolutional layers and k being an odd number, the overall receptive field size of each neuron in the output layer remains unchanged. As we shall see in experiments, the branched structure works much better than the original structure, and SORT consistently boosts the recognition performance beyond the improved baseline.

Another straightforward application of SORT lies in the family of deep residual networks [13]. Note that residual networks are already equipped with two-branch structures, *i.e.*, the input signal \mathbf{x} is followed by an identity mapping and the neural response after two convolutions. As a direct variant of (1), SORT modifies the original fusion function from $\mathbf{y}^R = \mathbf{x} + \mathbf{F}(\mathbf{x})$ to $\mathbf{y}^S = \mathbf{x} + \mathbf{F}(\mathbf{x}) + \mathbf{x} \odot \mathbf{F}(\mathbf{x})$. Note that in the residual networks, elements in either \mathbf{x} or $\mathbf{F}(\mathbf{x})$ may be negative [14], and we perform a ReLU activation on it before computing the

product term. Thus, the exact form of SORT in this case is $\mathbf{y}^S = \mathbf{x} + \mathbf{F}(\mathbf{x}) + \sigma[\mathbf{x}] \odot \sigma[\mathbf{F}(\mathbf{x})]$. Similarly, SORT does not change the receptive field size of an output neuron.

3.1.1 Implementation Details

In a two-branch network, each convolutional block actually contains three consecutive operations, *i.e.*, convolution, batch normalization [19] and ReLU activation [37]. Thus, both $\mathbf{F}_1(\mathbf{x})$ and $\mathbf{F}_2(\mathbf{x})$ are non-negative vectors. We simply compute the element-wise product of these two data blocks as in (1).

In a residual network, according to the implementation of [14], both \mathbf{x} and $\mathbf{F}(\mathbf{x})$ may contain negative entries. To avoid a positive response generated by the product of two negative responses, we feed both \mathbf{x} and $\mathbf{F}(\mathbf{x})$ into ReLU activation [37]. After computing the element-wise product, we perform element-wise square root for normalization [49]. To avoid numerical instability near the value 0, the square root function is defined as $\sqrt{u + \varepsilon}$, where $\varepsilon = 0.01$ is a small floating point number to prevent the gradient from being too large.

3.2. Cross-Branch Response Propagation

We first discuss the second-order term. According to our implementation, all the numbers fed into element-wise product are non-negative, *i.e.*, $\forall i, F_{1,i}(\mathbf{x}) \geq 0$ and $F_{2,i}(\mathbf{x}) \geq 0$. Therefore, the second-order term is either 0 or a positive value (when both $F_{1,i}(\mathbf{x})$ and $F_{2,i}(\mathbf{x})$ are positive). Consider two input pairs, *i.e.*, $(F_{1,i}(\mathbf{x}), F_{2,i}(\mathbf{x})) = (a, 0)$ or $(F_{1,i}(\mathbf{x}), F_{2,i}(\mathbf{x})) = (a_1, a_2)$ where $a_1 + a_2 = a$. In the former case we have $y_i^S = a$, but in the latter case we have $y_i^S = a + a_1 \times a_2$. The extra term, *i.e.*, $a_1 \times a_2$, is large when a_1 and a_2 are close, *i.e.*, $|a_1 - a_2|$ is small. We explain it as facilitating the *consistent* responses, *i.e.*, we reward the indices on which two branches have similar response values.

We also note that SORT leads to an improved way of gradient back-propagation. Since there exists a dyadic term $\mathbf{F}_1(\mathbf{x}; \boldsymbol{\theta}_1) \odot \mathbf{F}_2(\mathbf{x}; \boldsymbol{\theta}_2)$, the gradient of \mathbf{y}^S with respect to either one in $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ is related to another. Thus, when the parameter $\boldsymbol{\theta}_1$ needs to be updated, the gradient $\frac{\partial L}{\partial \boldsymbol{\theta}_1}$ is directly related to $\mathbf{F}_2(\mathbf{x})$:

$$\frac{\partial L}{\partial \boldsymbol{\theta}_1} = \left(\frac{\partial L}{\partial \mathbf{y}^S} \right)^\top \cdot [1 + \mathbf{F}_2(\mathbf{x}; \boldsymbol{\theta}_2)]^\top \cdot \frac{\partial \mathbf{F}_1(\mathbf{x}; \boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_1}, \quad (2)$$

and similarly, $\frac{\partial L}{\partial \boldsymbol{\theta}_2}$ is directly related to $\mathbf{F}_1(\mathbf{x})$. This prevents the gradients from being shattered as the network goes deep [1], and reduces the risk of structural over-fitting (*i.e.*, caused by the increasing number of network layers). As an example, we train deep residual networks [13] with different numbers of layers on the SVHN dataset [38], a relatively simple dataset for street house number recognition.

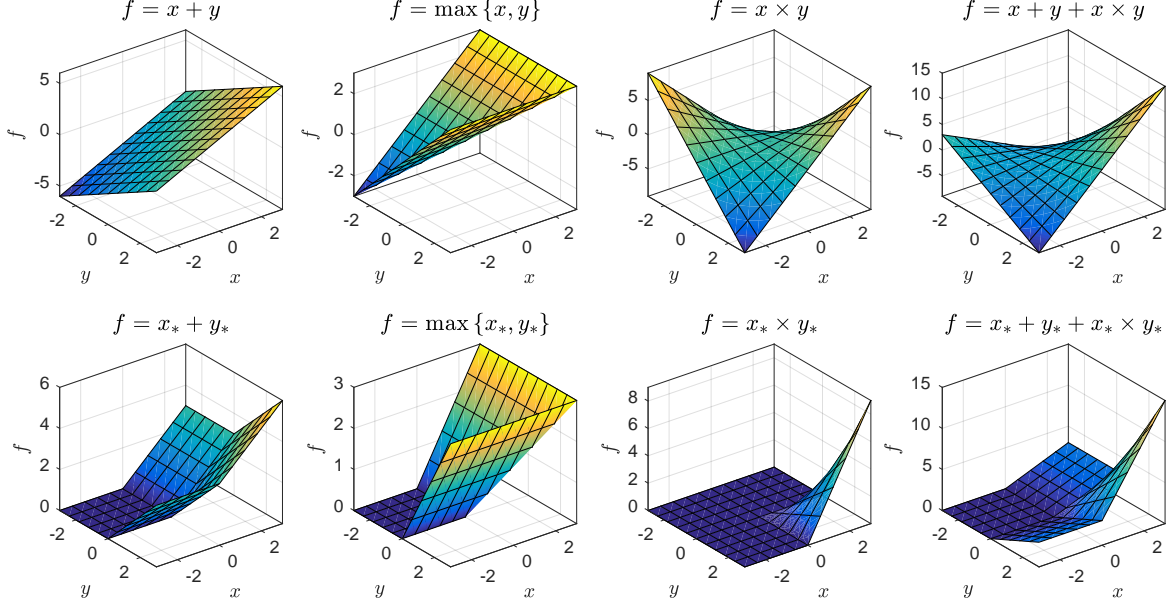


Figure 2. Comparison of different response transform functions. The second-order operation produces nonlinearity in a 2D space, while ReLU and max functions are piecewise linear. Here, $x_* \doteq \max\{x, 0\}$ and $y_* \doteq \max\{y, 0\}$.

Detailed experimental settings are illustrated in Section 4.1. The baseline recognition errors are 2.30% and 2.49% for the 20-layer and 56-layer networks, respectively, while these numbers become 2.26% and 2.19% after SORT is applied. SORT consistently improves the recognition rate, and the gain becomes more significant when a deeper network architecture is used.

In summary, SORT allows the network to consider cross-branch information in both forward-propagation and back-propagation. This strategy improves the reliability of neural responses, as well as the numerical stability in gradient computation.

3.3. Global Network Nonlinearity

Nonlinearity makes the major contribution to the representation ability of deep neural networks [21]. State-of-the-art networks are often equipped with sigmoid or ReLU activation [37] and/or max-pooling layers, and we argue that the proposed second-order term is a better choice. To compare these nonlinear operators, we plot these function graphs of different combinations in Figure 2. These functions, besides $f(x, y) = x + y$, bring in nonlinearity to some extent. If no second-order term is involved, the function $f(x, y)$ is piecewise linear, which means that nonlinearity only appears in a 1D subspace of the 2D space \mathbb{R}^2 . By adding the second-order term, nonlinearity exists in either \mathbb{R}^2 or $\mathbb{R}_*^2 = [0, +\infty)^2$.

Summarizing the cues above (cross-branch propagation and nonlinearity) leads to adding a second-order term which

+	max	\odot	LeNet	BigNet	ResNet
✓			11.10	6.86	7.60
	✓		11.07	7.01	7.55
		✓	11.03	—	—
✓	✓		11.02	6.90	7.63
✓		✓	10.34	6.60	7.14
	✓	✓	10.39	6.57	7.44
✓	✓	✓	10.80	6.65	7.90

Table 1. Recognition error rate (%) on the **CIFAR10** dataset with different fusion strategies. Here, +, max and \odot denote three dyadic operators, and multiple checkmarks in one row means to sum up the results produced by the corresponding operators. Sometimes, using the second-order terms alone results in non-convergence (denoted by —). All these numbers are the average over 3 individual runs, and the standard deviations are often 0.04%–0.08%.

involves neural responses from both branches. Hence, $\mathbf{F}_1 \odot \mathbf{F}_2$ is a straightforward and simple choice. We point out that an alternative choice of second-term nonlinearity is the square term, *i.e.*, $\mathbf{F}_1^2(\mathbf{x})$, where \cdot^2 denotes the element-wise operation. but we do not suggest this option, since this does not allow cross-branch response propagation. As a side note, an element-wise product term behaves similarly to a logical-and term, which is verified effective in learning feature representations in neural networks [35].

We experimentally verify the effectiveness of the term $\mathbf{F}_1 \odot \mathbf{F}_2$. We consider three types of fusion strategies, *i.e.*, $\mathbf{F}_1(\mathbf{x}) + \mathbf{F}_2(\mathbf{x})$, $\max\{\mathbf{F}_1(\mathbf{x}), \mathbf{F}_2(\mathbf{x})\}$ and $\mathbf{F}_1(\mathbf{x}) \odot \mathbf{F}_2(\mathbf{x})$.

To compare their performance, we apply different fusion strategies on different networks, and evaluate them on the **CIFAR10** dataset (detailed settings are elaborated in Section 4.1). Various combinations lead to different recognition results, which are summarized in Table 1.

We first note that the second-order operator \odot shall not be used alone, since this often leads to non-convergence especially in those very deep networks, *e.g.*, **BigNet** (19 layers) and **ResNet** (20 layers). The learning curves in Figure 3 also provide evidences to this point. It is well acknowledged that first-order terms are able to provide numerical stability, and help the training process converge [37] compared to some saturable activation functions such as sigmoid. On the other hand, when the second-order term is appended to either $+$ or \max , the recognition error is significantly decreased, which suggests that adding higher-order terms indeed increases the network representation ability, which helps to better depict the complicated feature space and achieve higher recognition rates. Missing either the first-order or second-order term harms the recognition accuracy of the deep network, thus we suggest to use a combination of linear and nonlinear terms in all the later experiments. In practice, we choose the linear sum mainly because it allows both branches to get trained in back-propagation, while the \max operator only updates half of the parameters at each time. In addition, the \max operator does not reward consistent responses as the second-order term does.

3.4. Relationship to Other Work

We note that some previous work also proposed to use a second-order term in network training. For example, the bilinear CNN [32] computes the outer-product of neural responses from two individual networks to capture feature co-occurrence at the same spatial positions. However, this operation often requires heavy time and memory overheads, as it largely increases the dimensionality of the feature vector, and consequently the number of trainable parameters. Training a bilinear CNN is often slow, even in the improved versions [9][30]. In comparison, the extra computation brought by SORT is merely ignorable ($< 5\%$).

In a spatial transformer network [20], the product operator is used to apply an affine transform on the neural responses. In some attention-based models [3][33], product operations are also used to adjust the intensity of neurons according to the spatial weights. We point out that SORT is generalized. Due to its simple mathematical form and efficient execution, it can be applied to many different network structures.

SORT is also related to the gating function used in recurrent neural network cells such as the long short-term memory (LSTM) [16] or the gated recurrent unit (GRU) [4]. There, element-wise product is used at each time step to regularize the memory cell and the hidden state. This operation

has also been explored in computer vision [44] to facilitate very deep network training. However, the focus of gating is to regularize activation and avoid gradient vanishing and/or exploding, while we use element-wise product to improve the representation ability of deep networks.

4. Experiments

We apply the second-order response transform (SORT) to several popular network architectures, including chain-styled networks (**LeNet**, **BigNet** and **AlexNet**) and two variants of deep residual networks. We verify significant accuracy gain over a wide range of visual recognition tasks.

4.1. Small-Scale Experiments

4.1.1 Settings

Three small-scale datasets are used in this section. Among them, the **CIFAR10** and **CIFAR100** datasets [23] are subsets drawn from the 80-million tiny image database [47]. Each set contains 50,000 training samples and 10,000 testing samples, and each sample is a 32×32 RGB image. In both datasets, training and testing samples are uniformly distributed over all the categories (**CIFAR10** contains 10 basic classes, and **CIFAR100** has 100 where the visual concepts are defined at a finer level). The **SVHN** dataset [38] is a larger collection for digit recognition, *i.e.*, there are 73,257 training samples, 26,032 testing samples, and 531,131 extra training samples. Each sample is also a 32×32 RGB image. We preprocess the data as in the previous literature [38], *i.e.*, selecting 400 samples per category from the training set as well as 200 samples per category from the extra set, using these 6,000 images for validation, and the remaining 598,388 images as training samples. We also use local contrast normalization (LCN) for data preprocessing [11].

Four baseline network architectures are evaluated.

- **LeNet** [26] is a relatively shallow network with 3 convolutional layers, 3 pooling layers and 2 fully-connected layers. All the convolutional layers have 5×5 kernels, and the input cube is zero-padded by a width of 2 so that the spatial resolution of the output remains unchanged. After each convolution including the first fully-connected layer, a nonlinear function known as ReLU [37] is used for activating the neural responses. This common protocol will be used in all the network structures. The pooling layers have 3×3 kernels, and a spatial stride of 2. We apply three training sections with learning rates of 10^{-2} , 10^{-3} and 10^{-4} , and 60K, 5K, and 5K iterations, respectively.
- A so-called **BigNet** is trained as a deeper chain-styled network. There are 10 convolutional layers, 3 pooling layers and 3 fully-connected layers in this architecture. The design of **BigNet** is similar to **VGGNet** [43], in

which small convolutional kernels (3×3) are used and the depth is increased. Following [36], we apply four training sections with learning rates of 10^{-1} , 10^{-2} , 10^{-3} and 10^{-4} , and 60K, 30K, 20K and 10K iterations, respectively.

- The deep residual network (**ResNet**) [13] brings significant performance boost beyond chain-styled networks. We follow the original work [13] to define network architectures with different numbers of layers, which are denoted as **ResNet-20**, **ResNet-32** and **ResNet-56**, respectively. These architectures differ from each other in the number of residual blocks used in each stage. Batch normalization is applied after each convolution to avoid numerical instability in this very deep network. Following the implementation of [55], we apply three training sections with learning rates of 10^{-1} , 10^{-2} , and 10^{-3} , and 32K, 16K and 16K iterations, respectively.
- The wide residual network (**WRN**) [56] takes the idea to increase the number of kernels in each layer and decrease the network depth at the same time. We apply the 28-layer architecture, denoted as **WRN-28**, which is verified effective in [56]. Following the same implementation of the original **ResNets**, we apply three training sections with learning rates of 10^{-1} , 10^{-2} and 10^{-3} , and 32K, 16K, and 16K iterations, respectively.

In all the networks, the mini-batch size is fixed as 100. Note that both **LeNet** and **BigNet** are chain-styled networks. Using the details illustrated in Section 3.1, we replace each convolutional layer using a two-branch, two-layer module with smaller kernels. This leads to deeper and more powerful networks, and we append an asterisk (*) after the original networks to denote them. SORT is applied to the modified network structure by appending element-wise product to linear sum.

4.1.2 Results

Results are summarized in Table 2. One can observe that SORT boosts the performance of all network architectures consistently. On both **LeNet** and **BigNet**, we observe significant accuracy gain brought by replacing of each convolutional layer as a two-branch module. SORT further improves recognition accuracy by using a more effective fusion function. In addition, we observe more significant accuracy gain when the network goes deeper. For example, on the 20-layer **ResNet**, the relative error rate drops are 4.79, 0.47% and 1.74% for **CIFAR10**, **CIFAR100** and **SVHN**, and these numbers become much bigger (12.70, 5.27% and 12.05%, respectively) on the 56-layer **ResNet**. This verifies our hypothesis in Section 3.2, that SORT alleviates the shattered gradient problem and helps training

Network	CF10	CF100	SVHN
Lee <i>et.al</i> [29]	7.97	34.57	1.92
Liang <i>et.al</i> [31]	7.09	31.75	1.77
Lee <i>et.al</i> [28]	6.05	32.37	1.69
Zagoruyko <i>et.al</i> [56]	5.37	24.53	1.85
Xie <i>et.al</i> [51]	5.31	25.01	1.67
Huang <i>et.al</i> [18]	5.25	24.98	1.75
Huang <i>et.al</i> [17]	3.74	19.25	1.59
LeNet	14.37	43.83	4.00
LeNet*	11.16	36.84	2.65
LeNet*-SORT	10.41	34.67	2.47
BigNet	7.55	30.47	2.21
BigNet*	6.92	29.43	2.17
BigNet*-SORT	6.81	28.10	2.12
ResNet-20	7.72	31.80	2.30
ResNet-20-SORT	7.35	31.65	2.26
ResNet-32	6.83	30.28	2.54
ResNet-32-SORT	6.33	29.61	2.22
ResNet-56	6.30	28.25	2.49
ResNet-56-SORT	5.50	26.76	2.19
WRN-28	4.81	21.90	1.93
WRN-28-SORT	4.48	21.52	1.48

Table 2. Recognition error rate (%) on small datasets and different network architectures. All the numbers are averaged over 3 individual runs, and the standard deviation is often less than 0.08%.

very deep networks more efficiently. Especially, based on **WRN-28**, one of the state-of-the-art structures, SORT reduces the recognition error rate of **SVHN** from 1.93% to 1.48%, giving a relatively 23.32% error drop, meanwhile achieving the new state-of-the-art (the previous record is 1.59% [17]). All these results suggest the usefulness of the second-order term in visual recognition.

4.1.3 Discussions

We plot the learning curves of several architectures in Figure 3. It is interesting to observe the convergence of network structures before and after using SORT. On the two-branch variants of both **LeNet** and **BigNet**, SORT allows each parameterized branch to update its weights based on the information of the other one, therefore it helps the network to get trained better (the testing curves are closer to 0). On the residual networks, as explained in Section 3.3, SORT introduces numerical instability and makes it more difficult for the network training to converge, thus in the first training section (*i.e.*, with the largest learning rate), the network with SORT often reports unstable loss values and recognition rates compared to the network without SORT. However, in the later sections, as the learning rate goes down and the training process becomes stable, the network with SORT

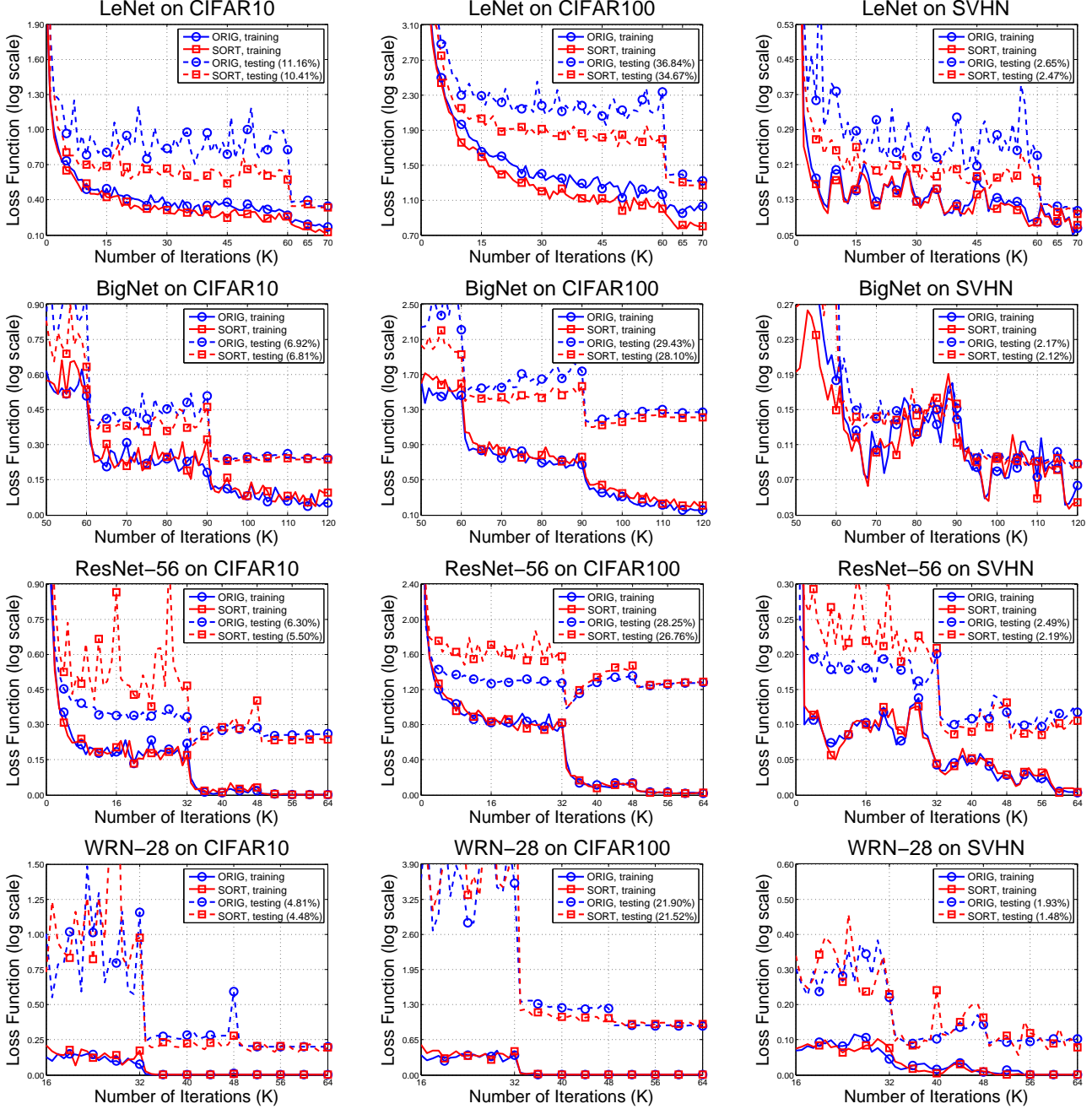


Figure 3. **CIFAR10**, **CIFAR100** and **SVHN** learning curves with different networks. Each number in parentheses denote the recognition error rate reported by the final model. Please zoom in for more details.

benefits from the increasing representation ability and thus works better than the baseline. As a side note, a similar loss value of SORT can lead to better recognition accuracy (*e.g.*, see the learning curves of **ResNet-56** and **WRN-28** on **CIFAR100**).

4.2. ImageNet Experiments

4.2.1 Settings

We further evaluate our approach on the **ILSVRC2012** dataset [42]. This is a subset of the **ImageNet** database [5] which contains 1,000 object categories. We train our models on the training set containing 1.3M images, and test them on the validation set containing 50K images. Two

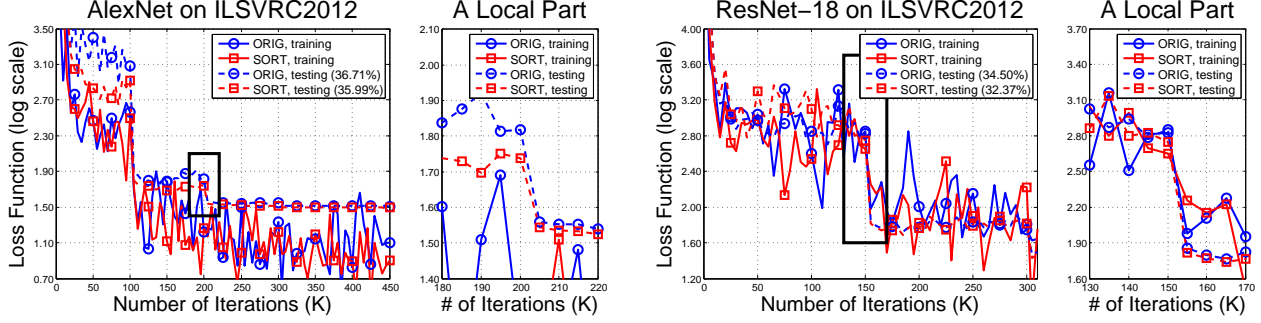


Figure 4. **ILSVRC2012** learning curves with **AlexNet** (left) and **ResNet-18** (right). Each number in parentheses denotes the top-1 error rate reported by the final model. For better visualization, we zoom in on a local part (marked by a black rectangle) of each learning curve.

Network	Top-1 Error	Top-5 Error
AlexNet	43.19	19.87
AlexNet*	36.71	14.77
AlexNet*-SORT	35.99	14.46
ResNet-18	34.50	13.33
ResNet-18-SORT	32.37	12.61

Table 3. Recognition error rate (%) on the **ILSVRC2012** dataset using different network architectures. All the results are reported using **one single crop** in testing.

Network	<i>pool-5</i>	<i>fc-6</i>	<i>fc-7</i>
AlexNet	69.19	71.51	69.47
(std deviation)	± 0.18	± 0.25	± 0.11
AlexNet*	74.20	76.54	74.42
(std deviation)	± 0.17	± 0.30	± 0.18
AlexNet*-SORT	74.88	77.12	75.06
(std deviation)	± 0.19	± 0.24	± 0.15

Table 4. Classification accuracy (%) on the **Caltech256** dataset using deep features extracted from different layers of different network structures.

network architectures are taken as the baseline. The first one is the **AlexNet** [24], a 8-layer network which is used for testing chain-styled architectures. As in the previous experiments, we replace each of the 5 convolutional kernels with a two-branch module, leading to a deeper and more powerful network structure, which is denoted as **AlexNet***. The second baseline is the 18-layer **ResNet** [13], which is the state-of-the-art network architecture for this large-scale visual recognition task. In both cases, we start from scratch, and train the networks with mini-batches of 256 images. The **AlexNet** is trained through 450K iterations, and the learning rate starts from 0.1 and drops by 1/10 after each 100K iterations. These numbers are 600K, 0.1 and 150K for the 18-layer **ResNet**, respectively.

4.2.2 Results

The recognition results are summarized in Table 3. All the numbers are reported by one single model. Based on the original chain-styled **AlexNet**, replacing each convolutional layer as a two-branch module produces 36.71% top-1 and 14.77% top-5 error rates, which is significantly lower than the original version, *i.e.*, 43.19% and 19.87%. This is mainly due to the increase in network depth. SORT further reduces the errors by 0.72% and 0.31 (or 1.96% and 2.10% relatively). On the 18-layer **ResNet**, the baseline top-1 and top-5 error rates are 34.50% and 13.33%, and SORT reduces them to 32.37% and 12.61% (6.17% and 5.71%

relative drop, respectively).

On a 4-GPU machine, **AlexNet*** and **ResNet-18** need an average of 10.5s and 19.3s to finish 20 iterations. After SORT is applied, these numbers becomes 10.7s and 19.9s, respectively. Given that only less than 5% extra time and no extra memory are used, we can claim the effectiveness and the efficiency of SORT in large-scale visual recognition.

4.2.3 Discussions

We also plot the learning curves of both architectures in Figure 4. Very similar phenomena are observed as in small-scale experiments. On **AlexNet*** which is the branched version of a chain-styled network, SORT helps the network to be trained better. Meanwhile, on **ResNet-18**, SORT makes the network more difficult to converge. But nevertheless, in either cases, SORT improves the representation ability and eventually helps the modified structure achieve better recognition performance.

4.3. Transfer Learning Experiments

We evaluate the transfer ability of the trained models by applying them to other image classification tasks. The **Caltech256** [12] dataset is used for generic image classification. We use the **AlexNet**-based models to extract from the *pool-5*, *fc-6* and *fc-7* layers, and adopt ReLU activation to filter out negative responses. The neural responses from the *pool-5* layer ($6 \times 6 \times 256$) are spatially averaged into a

256-dimensional vector, while the other two layers directly produce 4,096-dimensional feature vectors. We perform square-root normalization followed by ℓ_2 normalization, and use **LIBLINEAR** [7] as an SVM implementation and set the slack variable $C = 10$. 60 images per category are left out for training the SVM model, and the remaining ones are used for testing. The average accuracy over all categories is reported. We run 10 individual training/testing splits and report the averaged accuracy as well as the standard deviation. Results are summarized in Table 4. One can observe that the improvement on **ILSVRC2012** brought by SORT is able to transfer to **Caltech256**.

5. Conclusions

In this paper, we propose Second-Order Response Transform (**SORT**), an extremely simple yet effective approach to improve the representation ability of deep neural networks. SORT summarizes two neural responses by considering both sum and product terms, which leads to efficient information propagation throughout the network and more powerful network nonlinearity. SORT can be applied to a wide range of modern convolutional neural networks, and produce consistent recognition accuracy gain on some popular benchmarks. We also verify the increasing effectiveness of SORT on very deep networks.

In the future, we will investigate the extension of this simple operation. It remains open problems that whether SORT can be applied to multi-branch networks [46][17], or whether even higher-order terms can be added to enhance the representation ability of deep neural networks.

Acknowledgements

We thank Weichao Qiu, Zhuotun Zhu and Siyuan Qiao for instructive discussions.

References

- [1] D. Balduzzi, M. Frean, L. Leary, J. Lewis, K. W.-D. Ma, and B. McWilliams. The Shattered Gradients Problem: If ResNets are the Answer, then What is the Question? *arXiv preprint arXiv:1702.08591*, 2017.
- [2] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *International Conference on Learning Representations*, 2015.
- [3] L. Chen, Y. Yang, J. Wang, W. Xu, and A. Yuille. Attention to Scale: Scale-Aware Semantic Image Segmentation. *Computer Vision and Pattern Recognition*, 2016.
- [4] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv preprint arXiv:1412.3555*, 2014.
- [5] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. *Computer Vision and Pattern Recognition*, 2009.
- [6] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. *International Conference on Machine Learning*, 2014.
- [7] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.
- [8] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.
- [9] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell. Compact Bilinear Pooling. *Computer Vision and Pattern Recognition*, 2016.
- [10] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *Computer Vision and Pattern Recognition*, 2014.
- [11] I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio. Maxout Networks. *International Conference on Machine Learning*, 2013.
- [12] G. Griffin, A. Holub, and P. Perona. Caltech-256 Object Category Dataset. *Technical Report: CNS-TR-2007-001*, 2007.
- [13] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. *Computer Vision and Pattern Recognition*, 2016.
- [14] K. He, X. Zhang, S. Ren, and J. Sun. Identity Mappings in Deep Residual Networks. *European Conference on Computer Vision*, 2016.
- [15] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving Neural Networks by Preventing Co-adaptation of Feature Detectors. *arXiv preprint arXiv:1207.0580*, 2012.
- [16] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.
- [17] G. Huang, Z. Liu, and K. Weinberger. Densely Connected Convolutional Networks. *arXiv preprint arXiv:1608.06993*, 2016.
- [18] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Weinberger. Deep Networks with Stochastic Depth. *European Conference on Computer Vision*, 2016.
- [19] S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *International Conference on Machine Learning*, 2015.
- [20] M. Jaderberg, K. Simonyan, and A. Zisserman. Spatial Transformer Networks. *Advances in Neural Information Processing Systems*, 2015.
- [21] K. Jarrett, K. Kavukcuoglu, Y. LeCun, et al. What is the Best Multi-Stage Architecture for Object Recognition? *International Conference on Computer Vision*, 2009.
- [22] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. *ACM International Conference on Multimedia*, 2014.
- [23] A. Krizhevsky and G. Hinton. Learning Multiple Layers of Features from Tiny Images. *Technical Report, University of Toronto*, 1(4):7, 2009.

- [24] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 2012.
- [25] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. *Computer Vision and Pattern Recognition*, 2006.
- [26] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [27] Y. LeCun, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Handwritten Digit Recognition with a Back-Propagation Network. *Advances in Neural Information Processing Systems*, 1990.
- [28] C. Lee, P. Gallagher, and Z. Tu. Generalizing Pooling Functions in Convolutional Neural Networks: Mixed, Gated, and Tree. *International Conference on Artificial Intelligence and Statistics*, 2016.
- [29] C. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-Supervised Nets. *International Conference on Artificial Intelligence and Statistics*, 2015.
- [30] Y. Li, N. Wang, J. Liu, and X. Hou. Factorized Bilinear Models for Image Recognition. *arXiv preprint arXiv:1611.05709*, 2016.
- [31] M. Liang and X. Hu. Recurrent Convolutional Neural Network for Object Recognition. *Computer Vision and Pattern Recognition*, 2015.
- [32] T. Lin, A. RoyChowdhury, and S. Maji. Bilinear CNN Models for Fine-Grained Visual Recognition. *International Conference on Computer Vision*, 2015.
- [33] C. Liu, J. Mao, F. Sha, and A. Yuille. Attention Correctness in Neural Image Captioning. *AAAI Conference on Artificial Intelligence*, 2017.
- [34] J. Long, E. Shelhamer, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. *Computer Vision and Pattern Recognition*, 2015.
- [35] Y. Mansour. An $O(n \log \log n)$ Learning Algorithm for DNF under the Uniform Distribution. *Journal of Computer and System Sciences*, 50(3):543–550, 1995.
- [36] Nagadomi. The Kaggle CIFAR10 Network. <https://github.com/nagadomi/kaggle-cifar10-torch7/>, 2014.
- [37] V. Nair and G. Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. *International Conference on Machine Learning*, 2010.
- [38] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Ng. Reading Digits in Natural Images with Unsupervised Feature Learning. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.
- [39] F. Perronnin, J. Sanchez, and T. Mensink. Improving the Fisher Kernel for Large-scale Image Classification. *European Conference on Computer Vision*, 2010.
- [40] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN Features off-the-shelf: an Astounding Baseline for Recognition. *Computer Vision and Pattern Recognition*, 2014.
- [41] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems*, 2015.
- [42] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, pages 1–42, 2015.
- [43] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learning Representations*, 2015.
- [44] R. Srivastava, K. Greff, and J. Schmidhuber. Highway Networks. *International Conference on Machine Learning*, 2015.
- [45] R. Srivastava, K. Greff, and J. Schmidhuber. Training Very Deep Networks. *Advances in Neural Information Processing Systems*, 2015.
- [46] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going Deeper with Convolutions. *Computer Vision and Pattern Recognition*, 2015.
- [47] A. Torralba, R. Fergus, and W. Freeman. 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970, 2008.
- [48] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-Constrained Linear Coding for Image Classification. *Computer Vision and Pattern Recognition*, 2010.
- [49] Y. Wang, L. Xie, Y. Zhang, W. Zhang, and A. Yuille. Deep Collaborative Learning for Visual Recognition. *arXiv preprint arXiv:1703.01229*, 2017.
- [50] L. Xie, R. Hong, B. Zhang, and Q. Tian. Image Classification and Retrieval are ONE. *International Conference on Multimedia Retrieval*, 2015.
- [51] L. Xie, Q. Tian, J. Flynn, J. Wang, and A. Yuille. Geometric Neural Phrase Pooling: Modeling the Spatial Co-occurrence of Neurons. *European Conference on Computer Vision*, 2016.
- [52] L. Xie, J. Wang, W. Lin, B. Zhang, and Q. Tian. Towards Reversal-Invariant Image Representation. *International Journal on Computer Vision*, 2016.
- [53] L. Xie, J. Wang, Z. Wei, M. Wang, and Q. Tian. DisturbLabel: Regularizing CNN on the Loss Layer. *Computer Vision and Pattern Recognition*, 2016.
- [54] S. Xie and Z. Tu. Holistically-Nested Edge Detection. *International Conference on Computer Vision*, 2015.
- [55] J. Xu. Residual Network Test. <https://github.com/twtyggqyy/resnet-cifar10>, 2016.
- [56] S. Zagoruyko and N. Komodakis. Wide Residual Networks. *arXiv preprint arXiv:1605.07146*, 2016.